Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 81 (2016) 250 – 257

Procedia Computer Science

5th Workshop on Spoken Language Technology for Under-resourced Languages, SLTU 2016, 9-12 May 2016, Yogyakarta, Indonesia

# A Study of Statistical Machine Translation Methods for Under Resourced Languages

Win Pa Pa[a], Ye Kyaw Thu[b,c,*], Andrew Finch[b], Eiichiro Sumita[b]

[a] *Natural Language Processing Lab., University of Computer Studies, Yangon, Myanmar*
[b] *Advanced Speech Translation Research and Development Promotion Center, National Institute of Information and Communications Technology, Kyoto, Japan*
[c] *Language and Speech Science Research Lab., Department of Applied Mathematics, Global Information and Telecommunication Institute, Waseda University, Tokyo, Japan*

**Abstract**

This paper contributes an empirical study of the application of five state-of-the-art machine translation to the translation of low-resource languages. The methods studied were phrase-based, hierarchical phrase-based, the operational sequence model, string-to-tree, tree-to-string statistical machine translation methods between English (en) and the under resourced languages Lao (la), Myanmar (mm), Thai (th) in both directions. The performance of the machine translation systems was automatically measured in terms of BLEU and RIBES for all experiments. Our main findings were that the phrase-based SMT method generally gave the highest BLEU scores. This was counter to expectations, and we believe indicates that this method may be more robust to limitations on the data set size. However, when evaluated with RIBES, the best scores came from methods other than phrase-based SMT, indicating that the other methods were able to handle the word re-ordering better even under the constraint of limited data. Our study achieved the highest reported results on the data sets for all translation language pairs.



Peer-review under responsibility of the Organizing Committee of SLTU 2016

*Keywords:* Machine translation; Under resourced languages; Phrase-based; Hierarchical Phrase-based; Operation Sequence Model; Syntax-based

## 1. Introduction

Our main motivation for this research is to investigate machine translation performance with the dominant statistical machine translation (SMT) approaches on under resourced languages. We chose a selection of under-resource languages comprised of Lao, Myanmar and Thai for our study. We did experiments

* Corresponding author.
*E-mail address:* wasedakuma@gmail.com



Peer-review under responsibility of the Organizing Committee of SLTU 2016
doi:10.1016/j.procs.2016.04.057

with baseline phrase-based (PBSMT) and other advanced techniques hierarchical phrase-based (HPBSMT), operation sequence model (OSM), syntax-based models (SYN) of string-to-tree (S2T) and tree-to-string (T2S) methods. We trained PBSMT, HPBSMT, OSM, S2T and T2S machine translation systems using the ASEAN-MT parallel corpus (http://www.aseanmt.org/index.php) for each language pair (en-la, la-en, en-mm, mm-en, en-th and th-en)[1]. In this paper, to the best of our knowledge, we contribute the first comparative study of the five SMT methods on low-resource languages.

## 2. Related work

### *2.1. Myanmar*

To date, there have been very few studies on the automatic translation of Myanmar language. Ye Kyaw Thu et al.[2] studied word segmentation in the context of statistical machine translation using 7 different schemes, including a proposed unsupervised segmentation approach which did not exceed the performance of the simpler maximum matching approach. They hypothesized that the cause was a lack of data. Most of the approaches have been rule based, in[3] a method for word to phrase re-ordering for Myanmar-English translation based on English grammar rules was proposed. Thin Thin Wai et al.[4] studied Myanmar word disambiguation for Myanmar-English MT. In[5], a Myanmar phrase translation model with morphological analysis for Myanmar to English translation. All previous research has been based on very small parallel corpora (the largest being 13,042 sentence pairs).

### *2.2. Thai*

The first rule based English-Thai MT system was created by NECTEC, Thailand[6]. The technology is based on an English-to-Japanese machine translation system developed by NEC Corporation, Japan. The system employs the word syntactic and semantic information expressed in grammatical rules and dictionaries to analyze a source language (English) sentence and consecutively generate the target language (Thai) sentence.

In[7] a large-scale translation system between Thai and English was studied and the main focus was on methods for text preprocessing such as normalization and sentence-breaking. They report BLEU scores of around 0.2 in both directions.

### *2.3. Lao*

A phoneme-based transfer method for Thai to Lao machine translation was proposed in[8]. The most probable sequence of phonemes is generated by probabilistic GLR (PGLR) and Thai-Lao phoneme conversion rules are applied to obtain the Lao pronunciation. Morphological generation is then applied to the output of generated sequence phonemes to get the Lao translation. The system was evaluated on 35,125 Thai words and the conversion accuracy was 76% (without using a dictionary).

## 3. Methodology

In this section, we describe the methodology used in the machine translation experiments for this paper.

### *3.1. Phrase-based statistical machine translation (PBSMT)*

A PBSMT translation model is based on phrasal units[9,10]. Here, a phrase is simply a contiguous sequence of words and generally, not a linguistically motivated phrase. A phrase-based translation model typically gives better translation performance than word-based models. We can describe a simple phrase-based translation model consisting of phrase-pair probabilities extracted from corpus and a basic reordering model, and an algorithm to extract the phrases to build a phrase-table[11]. Figure 1 shows an example translation

process of the English sentence “I will meet my old friends this evening” into Myanmar with a PBSMT model.

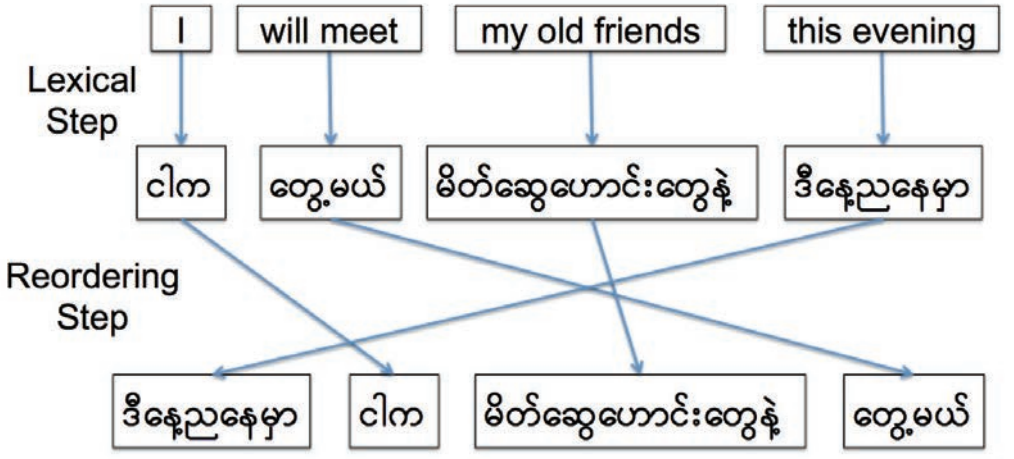



Fig. 1: Phrase-based translation

*3.2. Hierarchical phrase-based statistical machine translation (HPBSMT)*

The hierarchical phrase-based SMT approach is a model[12] based on synchronous context-free grammar. The model is able to be learned from a corpus of unannotated parallel text. The advantage this technique offers over the phrase-based approach is that the hierarchical structure is able to represent the word re-ordering process. The re-ordering is represented explicitly rather than encoded into a lexicalized re-ordering model (commonly used in purely phrase-based approaches). This makes the approach particularly applicable to language pairs that require long-distance re-ordering during the translation process[13]. An example of hierarchical phrase-based grammar between English and Myanmar is shown in Figure 2.

best for [X] ||| အတွက် အကောင်းဆုံး [X]
best for [X] ||| အတွက် အကောင်းဆုံး လို့ [X]
best for [X][X] [X] ||| [X][X] အတွက် အကောင်းဆုံး [X]
best for [X][X] [X] ||| [X][X] အတွက် အကောင်းဆုံး လို့ [X]
best for our family [X] ||| ကျွန်တော်တို့ မိသားစု အတွက် အကောင်းဆုံး [X]

Fig. 2: Hierarchical phrase-based translation

*3.3. Operation Sequence Model (OSM)*

The Operation Sequence Model (OSM)[14], combines the benefits of phrase-based and N-gram-based SMT[15] and remedies their drawbacks. It is based on minimal translation units, capture source and target context across phrasal boundaries and simultaneously generate source and target units. Providing a strong coupling of lexical generation and reordering gives a better reordering mechanism than PBSMT. The list of operations can be divided into two groups and they are five translation operations (Generate (X,Y), Continue Source Cept, Generate Identical, Generate Source Only (X) and Generate Target Only (Y)) and

three reordering operations (Insert Gap, Jump Back (N) and JumpForward). Figure 3 shows an example translation process of English sentence "I live in Bali" into Myanmar with the OSM.

### *3.4. Syntax-based machine translation (S2T and T2S)*

Syntax-based machine translation models use a grammar consisting of SCFG (Synchronous Context-Free Grammar) rules with syntactic labels `http://www.statmt.org/moses/?n=Moses.SyntaxTutorial`. To get the annotation, a syntactic parser is required. Syntactic labels provide the structure of a sentence and can also indicate relationships to structural units in other languages. S2T translation exploits target-side syntax[16], while T2S translation, source-side syntactic tree annotation is employed[17 18]. For example, the S2T model provides a method to transduce a negative verb string မ VB ဘူး in the source language of Myanmar into a structural representation in the target language of English (see Figure 4a). Conversely, a T2S model provides method to transduce structural representations in the source language English sentence "The girl ate the cake yesterday" into a string in the target language (see Figure 4b).

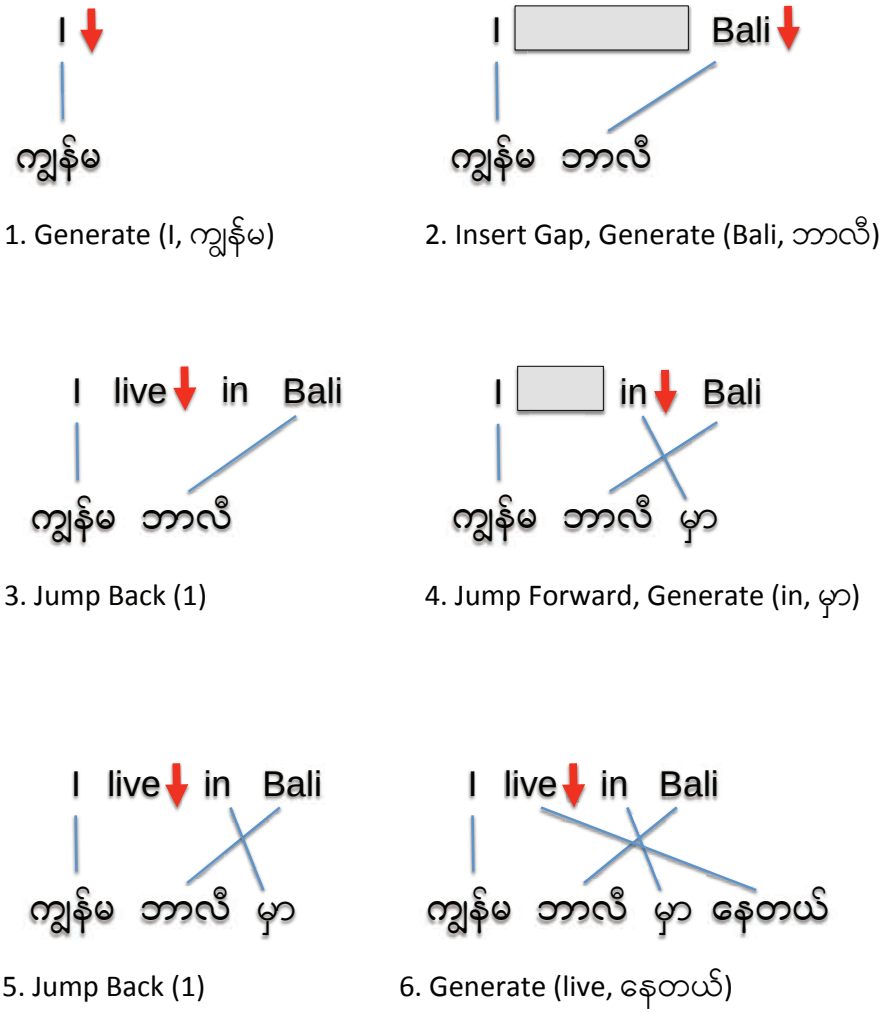
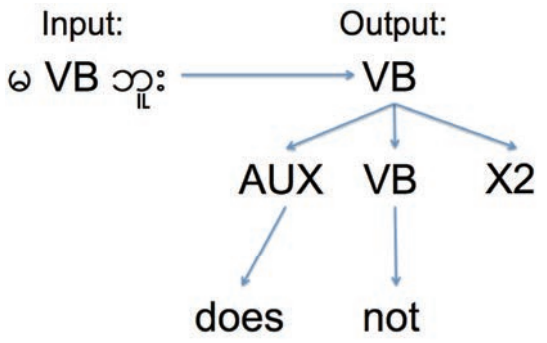
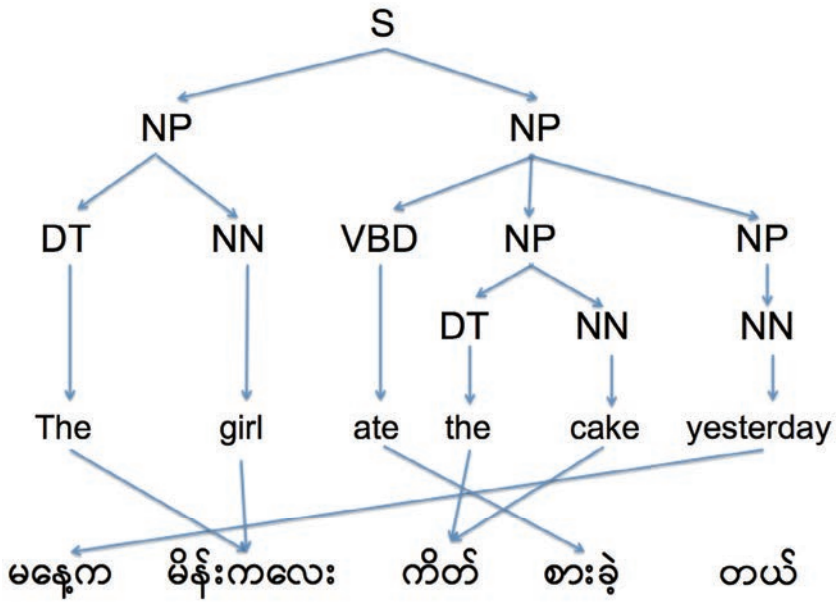



Fig. 3: Operation sequence translation




(a) String-to-Tree translation.




(b) Tree-to-String translation.

Fig. 4: Syntax based translation

## 4. Experiments

### *4.1. Corpus statistics*

We used four languages from the ASEAN-MT Parallel Corpus[1] without name entity tags, which is a parallel corpus in the travel domain. It contains six main categories and they are people (greeting, introduction and communication), survival (transportation, accommodation and finance), food (food, Beverage and restaurant), fun (recreation, traveling, shopping and nightlife), resource (number, time and accuracy), special needs (emergency and health). The languages were en, la, mm and th. 20,000 sentences were used for training, 500 sentences for development and 300 sentences for evaluation.

### *4.2. Moses SMT system*

We used the PBSMT, HPBSMT, OSM, S2T and T2S system provided by the Moses toolkit [19] for training the PBSMT, HPBSMT, OSM, S2T and T2S statistical machine translation systems. The word segmented source language was aligned with the word segmented target languages using GIZA++ [20]. The alignment was symmetrized by grow-diag-final-and heuristic [21]. The lexicalized reordering model was trained with the msd-bidirectional-fe option [22]. We use SRILM for training the 5-gram language model with interpolated modified Kneser-Ney discounting [23,24]. Minimum error rate training (MERT) [25] was used to tune the decoder parameters and the decoding was done using the Moses decoder (version 2.1.1) [19]. We used default settings of Moses for all experiments. We used the Berkeley Parser [26] for tree annotation of English for S2T and T2S experiments. According to our knowledge, there is no publicly available tree parser for Lao, Myanmar and Thai languages and thus annotated tree is used only for English language for S2T and T2S experiments.

## 5. Evaluation

We used two automatic criteria for the evaluation of the machine translation output. One was the de facto standard automatic evaluation metric Bilingual Evaluation Understudy (BLEU) [27] and the other was the Rank-based Intuitive Bilingual Evaluation Measure (RIBES) [28]. The BLEU score measures the precision of $n$-grams (over all $n \leq 4$ in our case) with respect to a reference translation with a penalty for short translations [27]. Intuitively, the BLEU score measures the adequacy of the translations and large BLEU scores are better. RIBES is an automatic evaluation metric based on rank correlation coefficients modified with precision and special care is paid to word order of the translation results. The RIBES score is suitable for distant language pairs such as Myanmar and English, Myanmar and Thai [28]. Large RIBES scores are better.

Table 1: BLEU and RIBES scores for PBSMT, HPBSMT, OSM, S2T and T2S.

| **Source-Target** | **BLEU Scores** | | | | **RIBES Scores** | | | |
|---|---|---|---|---|---|---|---|---|
| | PBSMT | HPBSMT | OSM | S2T or T2S | PBSMT | HPBSMT | OSM | S2T or T2S |
| **en-la** | **20.87** | 18.94 | 19.84 | 15.88 | 37.26 | 36.82 | **37.67** | 36.40 |
| **la-en** | **31.41** | 30.73 | 31.35 | 23.67 | 66.54 | **67.37** | 65.33 | 20.56 |
| **en-mm** | 10.71 | **12.53** | 10.22 | 9.20 | 58.24 | **62.72** | 59.79 | 56.87 |
| **mm-en** | **21.65** | 20.95 | 21.10 | 15.22 | 66.12 | 65.93 | **65.98** | 61.12 |
| **en-th** | 37.33 | **38.60** | 36.91 | 36.23 | 78.04 | 78.84 | 78.26 | **79.73** |
| **th-en** | **36.98** | 35.45 | 36.74 | 26.14 | **83.07** | 82.43 | 82.86 | 71.65 |

## 6. Results and discussion

The BLEU and RIBES score results for machine translation experiments with PBSMT, HPBSMT, OSM, S2T or T2S are shown in Table 1. Bold numbers indicate the highest scores of the different approaches. Comparing to existing published (`http://www.aseanmt.org/index.php?q=index/status_update`) baselines of PBSMT and HPBSMT from the Network-based ASEAN Languages Translation Public Service, the configurations used in our experiments achieved higher scores for all approaches.

Table 1) gives the BLEU and RIBES scores for all systems. Although PBSMT approach gave rise to the highest BLEU scores for most language pairs, the highest RIBES scores are came from the HPBSMT, OSM, S2T and T2S approaches.

## 7. Conclusion

This paper has presented the first comparative study of five major machine translation approaches applied to low-resource languages. We studied the application of PBSMT, HPBSMT, T2S, S2T and OSM translation methods to the translation of limited quantities of travel domain data between English and {Thai, Laos, Myanmar} in both directions.

Our experimental results indicate that in terms of adequacy (as measured by BLEU score), the PBSMT approach produced the highest quality translations. This was counter to expectations for some of the language pairs that would require long distance re-ordering during the translation process, thereby favoring the other methods. In terms of word order (as measured by the RIBES score), the PBSMT fared the worst. It was not possible to determine the best method here from the other methods. Therefore we provisionally conclude that overall, the simpler PBMSMT method seems more robust to training on very limited amounts of data, but that it still has issues with word order. However, automatic metrics can sometimes be misleading and believe a future human evaluation with bilingual judges[29] would be required to gain a more complete understanding of the relative merits of the machine translation approaches we studied when applied to these low-resource languages.

## Acknowledgements

Thanks to National Electronics and Computer Technology Center (NECTEC), Thailand for sharing ASEAN-MT Corpus.

## References

1. Prachya, B., Thepchai, S.. Technical report for the network-based asean language translation public service project. In: *Online Materials of Network-based ASEAN Languages Translation Public Service for Members*. NECTEC; 2013, .
2. Thu, Y.K., Finch, A., Sagisaka, Y., Sumita, E.. A study of myanmar word segmentation schemes for statistical machine translation. *Proceeding of the 11th International Conference on Computer Applications* 2013;:167–179.
3. Win, A.T.. Words to phrase reordering machine translation system in myanmar-english using english grammar rules. 2011. doi:10.1109/ICCRD.2011.5764243.
4. Wai, T.T., Htwe, T.M., Thein, N.L.. Article: Automatic reordering rule generation and application of reordering rules in stochastic reordering model for english-myanmar machine translation. *International Journal of Computer Applications* 2011;**27**(8):19–25. Full text available.
5. Zin, T.T., Soe, K.M., Thein, N.L.. *Translation model of Myanmar phrases for statistical machine translation.* Berlin: Springer. ISBN 978-3-642-25943-2; 2012, p. 235–242. doi:10.1007/978-3-642-25944-9_31.
6. Virach, S., Paisarn, C., Monthika, B.. Article: Parsit: Online english-thai machine translation service. *Language Issues in Digital Publishing, Asian/Pacific Book Development (ABD)* 2001;**31**(3):6–7.
7. Slayden, G., Hwang, M.Y., Schwartz, L.. Large-scale thai statistical machine translation. Tech. Rep. MSR-TR-2010-41; 2010. URL: http://research.microsoft.com/apps/pubs/default.aspx?id=130868.
8. Virach, S., Chumpol, M.. Thai-lao machine translation based on phoneme transfer. In: *In Proceedings of the 14th IPSJ National Conference*. 2008, p. 65–68.
9. Koehn, P., Och, F.J., Marcu, D.. Statistical phrase-based translation. In: *HLT-NAACL*. 2003, URL: http://acl.ldc.upenn.edu/N/N03/N03-1017.pdf.
10. Och, F.J., Marcu, D.. Statistical phrase-based translation. 2003, p. 127–133.
11. Specia, L.. Tutorial, fundamental and new approaches to statistical machine translation. In: *International Conference Recent Advances in Natural Language Processing*. 2011, .
12. Chiang, D.. Hierarchical phrase-based translation. *Comput Linguist* 2007;**33**(2):201–228. URL: http://dx.doi.org/10.1162/coli.2007.33.2.201. doi:10.1162/coli.2007.33.2.201.
13. Braune, F., Gojun, A., Fraser, A.. Long-distance reordering during search for hierarchical phrase-based smt. In: *EAMT 2012: Proceedings of the 16th Annual Conference of the European Association for Machine Translation, Trento, Italy*. Citeseer; 2012, p. 177–184.
14. Durrani, N., Schmid, H., Fraser, A.M.. A joint sequence translation model with integrated reordering. In: Lin, D., Matsumoto, Y., Mihalcea, R., editors. *ACL*. The Association for Computer Linguistics. ISBN 978-1-932432-87-9; 2011, p. 1045–1054. URL: http://dblp.uni-trier.de/db/conf/acl/acl2011.html#DurraniSF11.
15. Mariòo, J.B., Banchs, R.E., Crego, J.M., de Gispert, A., Lambert, P., Fonollosa, J.A.R., et al. N-gram-based machine translation. *Comput Linguist* 2006;**32**(4):527–549. URL: http://dx.doi.org/10.1162/coli.2006.32.4.527. doi:10.1162/coli.2006.32.4.527.

16. Zollmann, A., Venugopal, A., Och, F.J., Ponte, J.M.. A systematic comparison of phrase-based, hierarchical and syntax-augmented statistical MT. In: *COLING 2008, 22nd International Conference on Computational Linguistics, Proceedings of the Conference, 18-22 August 2008, Manchester, UK*. 2008, p. 1145–1152. URL: http://www.aclweb.org/anthology/C08-1144.
17. Huang, L., Knight, K., Joshi, A.. A syntax-directed translator with extended domain of locality. In: *Proceedings of the Workshop on Computationally Hard Problems and Joint Inference in Speech and Language Processing*; CHSLP '06. Stroudsburg, PA, USA: Association for Computational Linguistics; 2006, p. 1–8. URL: http://dl.acm.org/citation.cfm?id=1631828.1631829.
18. Hopkins, M., Kuhn, J.. Machine translation as tree labeling. In: *Proceedings of the NAACL-HLT 2007/AMTA Workshop on Syntax and Structure in Statistical Translation*; SSST '07. Stroudsburg, PA, USA: Association for Computational Linguistics; 2007, p. 41–48. URL: http://dl.acm.org/citation.cfm?id=1626281.1626287.
19. Koehn, P., Haddow, B.. Edinburgh's Submission to all Tracks of the WMT2009 Shared Task with Reordering and Speed Improvements to Moses. In: *Proceedings of the Fourth Workshop on Statistical Machine Translation*. 2009, p. 160–164.
20. Och, F.J., Ney, H.. Improved statistical alignment models. In: *ACL00*. Hong Kong, China; 2000, p. 440–447.
21. Koehn, P., Och, F.J.,, , Marcu, D.. Statistical phrase-based translation. In: *In Proceedings of the Human Language Technology Conference*. Edmonton, Canada; 2003, .
22. Tillmann, C.. A unigram orientation model for statistical machine translation. In: *Proceedings of HLT-NAACL 2004: Short Papers*; HLT-NAACL-Short '04. Stroudsburg, PA, USA: Association for Computational Linguistics. ISBN 1-932432-24-8; 2004, p. 101–104. URL: http://dl.acm.org/citation.cfm?id=1613984.1614010.
23. Stolcke, A.. SRILM - An Extensible Language Modeling Toolkit. In: *Proceedings of the International Conference on Spoken Language Processing*; vol. 2. Denver; 2002, p. 901–904.
24. Chen, S.F., Goodman, J.. An empirical study of smoothing techniques for language modeling. In: *Proceedings of the 34th annual meeting on Association for Computational Linguistics*. Association for Computational Linguistics; 1996, p. 310–318.
25. Och, F.J.. Minimum error rate training for statistical machine translation. In: *Proceedings of the 41st Meeting of the Association for Computational Linguistics (ACL 2003)*. Sapporo, Japan; 2003, .
26. Petrov, S., Barrett, L., Thibaux, R., Klein, D.. Learning accurate, compact, and interpretable tree annotation. In: *ACL 2006, 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference, Sydney, Australia, 17-21 July 2006*. 2006, URL: http://aclweb.org/anthology/P06-1055.
27. Papineni, K., Roukos, S., Ward, T., Zhu, W.. *Bleu: a Method for Automatic Evaluation of Machine Translation*. Thomas J. Watson Research Center: IBM Research Report rc22176 (w0109022); 2001.
28. Isozaki, H., Hirao, T., Duh, K., Sudoh, K., Tsukada, H.. Automatic evaluation of translation quality for distant language pairs. In: *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*; EMNLP '10. Stroudsburg, PA, USA: Association for Computational Linguistics; 2010, p. 944–952. URL: http://dl.acm.org/citation.cfm?id=1870658.1870750.
29. Vilar, D., Leusch, G., Ney, H., Banchs, R.E.. Human evaluation of machine translation through binary system comparisons. In: *Second Workshop on Statistical Machine Translation*. Prague, Czech Republic: Association for Computational Linguistics; 2007, p. 96–103. URL: http://www.aclweb.org/anthology/W/W07/W07-0713.